原著論文

２標本の位置の検定に関するコンピュータ・シミュレーション（II）

天野弘美[1),2)]、近藤雅人[2)]、小倉　浩[2)]、高木利一[2)]

[1)] 昭和大学大学院保健医療学研究科

[2)] 昭和大学富士吉田教育部

要　旨

２標本の分布位置に関する検定についてのコンピュータ・シミュレーションを実行し、適切な検定手法をいかに選択するかという問題について調べた。今回の報告では、特に実用的に広く用いられている多段階検定についての検証を行った。シミュレーションでは、平均が等しく分散が異なる二つの正規分布を母集団分布として標本を発生させた場合について多段階検定を実行し、第1種の過誤の発生率を評価した。それぞれの母集団の分散の設定値および発生させた標本サイズによっては、第１種の過誤の発生率が増加し、多段階検定を行わずに最初から Welch 検定を適用した場合のほうが第１種の過誤の発生率を低く抑えることが可能になることが明らかになった。

Key Words：２標本の位置の検定、多段階検定、t-検定、Welch 検定、Wilcoxon の順位和検定

緒　言

２標本の位置に関する検定は、医療研究において頻出する代表的な検定の１つである[1), 2)]。この検定を行う際に最も大切なことは、データの母集団分布に応じて最も適切な検定手法を選択することで、以下の２点に特に留意する必要がある。

１　２標本の位置に関する各検定手法は、母集団分布が正規性および等分散性を満たすか否かによって、その適用条件がそれぞれ異なる[3)]。

２　医療データにおいては、母集団の正規性および等分散性が仮定できない例が頻出する[1), 2)]。

以下に上記２について、２つの具体例を挙げる。１つは、ホルモン療法の血栓塞栓症への影響を調べた事例である。この事例は、治療開始６週間後の高投与量患者と低投与量患者それぞれに対して、フィブリン生成のマーカーとしての D-ダイマーの分布を調べたものである[4), 5)]。この分布は、低投与量患者では非常に歪んだ分布となり、また高投与量患者に対してはやや歪んだ分布となることが明らかとなっている。すなわち、この例では母集団はどちらも正規分布に従っているとは考えられず、また等分散性も（厳密には）仮定できないということになる。

もう一つの事例は、化学療法後の悪性リンパ腫患者に対して末梢血前駆細胞を移植し、好中球の回復度を測定した事例である[6), 7)]。測定は、ホジキンリンパ腫とそれ以外のリンパ腫の２つの群に分けて行われ、その結果を比較するという研究である。この研究事例では、母集団の正規性および等分散性の確認について一応の考慮が行われているが、データ変換によりこの問題に対処すべきであるとの誤った解決法が示されている[6)]。

このように、医療研究において行われる標本の位置に関する検定の多くは、最も適切な検定手法を選択するという視点を欠いたものであることが報告されている[5), 7)]。この視点がなければ、測定された医療データが正しいものであったとしても、誤った結論が導かれる可能性が大きい。すなわち、適切な検

定手法を選択するための明確な基準を明らかにすることは、これからの医療研究の進展に不可欠であると言える。

前報[3]では、最適な検定手法を選択するための指針の一つとして、２標本の位置に関する代表的な５つの検定手法について、個々の検定手法の安定性を調べた。しかし、検定結果の信頼性に影響を与える要因は個々の検定手法の安定性だけではない。これ以外の要因として、現在広く使用されている多段階検定が相応の妥当性を有しているかという問題は特に重要で、この検証がなければ最終結果の信頼性は大きく損なわれてしまう。

そこで、本研究では多段階検定そのものの妥当性を検証するためのコンピュータ・シミュレーションを行う。ここで、多段階検定とは２標本の母集団分布のそれぞれが正規分布に従っているか否かを判断するための正規性の検定、および２標本の母分散が等しいと見なせるかどうかの等分散検定を、得られている２つの標本に対してまず行い、これらの検定結果によって複数存在する２標本の位置に関する検定手法の中から適切なものを選択するという枠組みであり、実用的に広く用いられているものである。この枠組みの問題点は、最終目的である２標本の位置に関する検定を行う前に、正規性の検定および等分散検定という２つの検定を経ることで、最終的な検定結果の信頼性が損なわれる可能性が無視できない点にある。多段階検定の枠組みを使用することの妥当性を調べ、もし問題があるとすれば多段階検定によってどのような影響がどの程度生じているのかを明らかにすることが本研究の目的である。

方　法

１．多段階検定の枠組み

多段階検定の枠組みを使用して、最終的な２標本の位置に関する検定を選択するための手順として、妥当と考えられるものを図１に示す。図１では、位置に対する検定手法として t-検定[8]、Welch 検定[9]、Wilcoxon 検定[10), 11)]、Brunner-Munzel 検定[12)]の４種類を使用する。それぞれの適用条件および帰無仮説を表１に示す。t-検定および Welch 検定については、図１に示した正規性の検定および等分散性検定の結果と、表１に示した適用条件とがそのまま対応していることが分かる。すなわち、正規性の検定、等分散性検定のいずれにおいても帰無仮説が採択される場合は２標本の位置に関する検定手法として t-検定が選択され、また正規性の検定では帰無仮説が採択され、等分散性検定では帰無仮説が棄却される場合は Welch 検定が選択される。また、Brunner-Munzel 検定の適用条件は（現実に得られる測定データを考える限り）特にないと言えるため、図１において正規性の検定および等分散検定のいずれの検定においてもそれぞれの帰無仮説が棄却された場合に、Brunner-Munzel 検定が選択されるようになっている。

t-検定、Welch 検定、Brunner-Munzel 検定については、図１に示した各検定手法の選択過程と表１に示した各検定手法の適用条件が一致していることを述べた。これに対して Wilcoxon の順位和検定については、図１における処理（Wilcoxon の順位和検定が選択される経緯）と表１における Wilcoxon の順位和検定の適用条件とは直接対応していない。表１における、Wilcoxon の順位和検定の適用条件 $f_A(X_A)=f_B(X_B+\Delta X)$ の意味は、２つの母集団分布の確率密度関数が平行移動により重ねることができるという意味である[13)]。しかし、限られたサイズの標本をもとに高次のモーメントまでの比較を行うことは現実的ではない。そこで実際の適用条件の確認の際には、２次のモーメントである分散が一致していれば、第１群と第２群の関数形は（平行移動の任意性を除けば）一致しているものと見なすという立場を採用することにする。これが、図１に示した Wilcoxon 検定に至る途中での等分散検定の意味である。

また、表１における Wilcoxon の順位和検定の帰無仮説は、上記適用条件が満たされている場合に、２つの分布関数を重ねるために必要な平行移動量が０であるという意味である。言い換えれば、分布の位置に関する代表値である平均値、中央値、最頻値などがそれぞれ一致するという意味である。

これまで述べてきた解釈から、図１に示した２標本の位置に関する検定手法の選択過程は、それなりの根拠がある、納得のできるものであることが分かる。しかしながら、現在多くの研究者が用いる検定手法の選択フローは図１の過程ではなく、図２の過

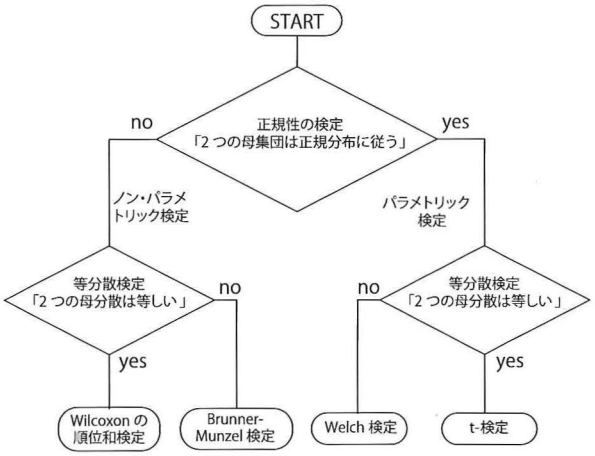

図１　多段階検定の枠組みによる、２標本の位置に関する検定の選択手順。現状で、最も妥当と考えられる手順を示した．yes は、対応する検定の帰無仮説が採択された場合を表し、逆に no は帰無仮説が棄却された場合を意味している。

表１　図１で使用されている４種類の検定手法の適用条件および帰無仮説のまとめ。それぞれの母集団は連続型確率分布に従うと仮定している。2つの母集団をそれぞれ母集団 A および母集団 B とし、また各母集団分布に対して、その平均，標準偏差，確率密度関数をそれぞれ $\mu_A, \sigma_A, f_A(X_A)$ および $\mu_B, \sigma_B, f_B(X_B)$ とした。Brunner-Munzel 検定および Wilcoxon の順位和検定の記述内容については，本文参照のこと。

検　定　手　法	適　用　条　件	帰　無　仮　説
t-検定	２つの母集団の正規性および $\sigma_A = \sigma_B$	$\mu_A = \mu_B$
Welch 検定	２つの母集団の正規性	$\mu_A = \mu_B$
Wilcoxon の順位和検定	$f_A(X_A) = f_B(X_B + \Delta X)$	$\Delta X = 0$
Brunner-Munzel 検定	特になし	$P(X_A > X_B) = 1/2$

程となっている。図１と比較すると、２標本の位置に関する検定手法として、Brunner-Munzel 検定が除外されていることが大きな違いである。

図２が多く使用されている理由として、Brunner-Munzel 検定は比較的最近開発された検定手法であり、多くの研究者はその存在を知らないことが挙げられる。

２．シミュレーション方法

本研究の主要目的は、多段階検定の妥当性を検証することである。そのためには、実際に行われている多段階検定の手順をコンピュータ・シミュレーションにより再現する必要がある。具体的な手順を以下に述べる。

1　使用する検定手法の帰無仮説が成立するように、２つの母集団分布の位置関係を決定する。
2　上記手順で決定した２つの各母集団分布から、第１群、第２群それぞれの標本を発生させる。
3　生成された２つの標本に対して、正規性の検定および等分散検定を実行し、その結果により使用する２標本の位置に関する検定の検定手法を選択する。
4　各標本に対して２標本の位置に関する検定を実行し、帰無仮説が棄却されたかをチェックする。
5　手順２～４を多数回繰り返し、第１種の過誤の発生率を求める。

図１の多段階検定の枠組みが図２と比較してより妥当であることはすでに述べたが、今回は図２の枠組みを利用してシミュレーションを行った。実際に現在多く使用されている枠組みが図２であることから、その妥当性を確認することが重要であると考えたからである。従って、本研究では図２の枠組みを使用して、２つの母集団として正規分布のみを使用し、その平均値を揃えた状態で標本を抽出することとする。

有意水準の設定　図２の枠組みで多段階検定を実行する場合には、３種類の有意水準の設定が必要となる。すなわち、正規性の検定で使用する有意水準 α_{nor}、等分散検定で使用する有意水準 α_σ およ

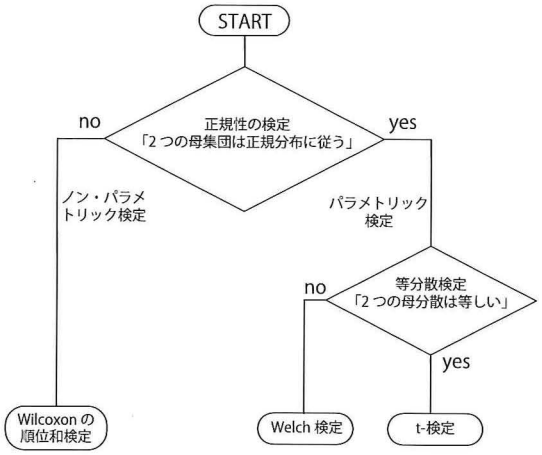

図２　多段階検定の枠組みによる２標本の位置に関する検定の選択手順。多くの研究者が従っていると思われる手順を示した。

び２標本の位置の検定（t-検定/Welch 検定/Wilcoxon 検定のいずれか）で使用する有意水準 α_μ である。シミュレーションでは、２標本の位置の検定に使用する有意水準はいずれの検定手法においても α_μ=0.05 を使用した。これに対して、正規性の検定および等分散検定に対する有意水準は $\alpha_{nor}=\alpha_\sigma$=0.05 とした場合と、$\alpha_{nor}=\alpha_\sigma$=0.2 とした場合の２通りについてシミュレーションを実行した。有意水準を0.2に設定することは通常は行わないが、今回の研究では以下を考慮して使用することとする。

* 多段階検定においては、正規性の検定における有意確率が通常設定される有意水準である0.05と比較して十分大きいことが、母集団分布を正規分布であると判断する場合の根拠として用いられる。十分大きい値の一例として 0.2 を使用する。
* 多段階検定においては、等分散検定における有意確率が通常設定される有意水準である 0.05 と比較して十分大きいことが、２つの母集団分布の分散が等しいと判断する場合の根拠として用いられる。十分大きい値の一例として 0.2 を使用する。

プログラム　使用したシミュレーションプログラムは、基本的には図２に示した枠組みに従って検定を多数回繰り返すプログラムである。すなわち、正規性の検定および等分散検定を経て、その結果により２標本の位置に関する検定手法を一つだけ選択して実行することを繰り返している。ただし、図２に示した枠組みの妥当性の検証のために、生成された２組の標本に対して、事前に正規性の検定および等分散検定を経ることなく無条件にt-検定、Welch 検定および Wilcoxon の順位和検定を実行する処理を追加してある。プログラム作成には統計解析環境 R[14]（Ver.2.12.0）を使用し、10万回の試行回数でシミュレーションを実行した。

結　果

シミュレーション結果を図３に示す。ここでは２つの母集団正規分布の平均をすべての場合に共通して0.0に設定した。図３(a)は、標本サイズの影響を調べるために第１群の標本サイズを10、20、30、40、50と変化させた場合の、第１種の過誤の発生率を示したものである。この際、第２群の標本サイズは10に固定し、また第１群の標準偏差は1.5に、第２群の標準偏差は1.0に固定した。多段階検定との比較のために、図３(a)には、正規性の検定および等分散検定を経ずに無条件で Welch 検定を適用した場合、無条件でt-検定を適用した場合および無条件で Wilcoxon 検定を適用した場合の第１種の過誤の発生率もあわせて示した。これら無条件 Welch 検定、無条件 t-検定および無条件 Wilcoxon 検定における有意水準は、いずれも0.05に設定した。

一方図３(b)は、標準偏差の不均一性の影響を調べるために第1群の標準偏差を0.5、1.0、1.5、2.0、2.5と変化させた場合の、第１種の過誤の発生率を示したものである。この際、第２群の標準偏差は1.0に固定し、また第１群の標本サイズは10に、第２群の標

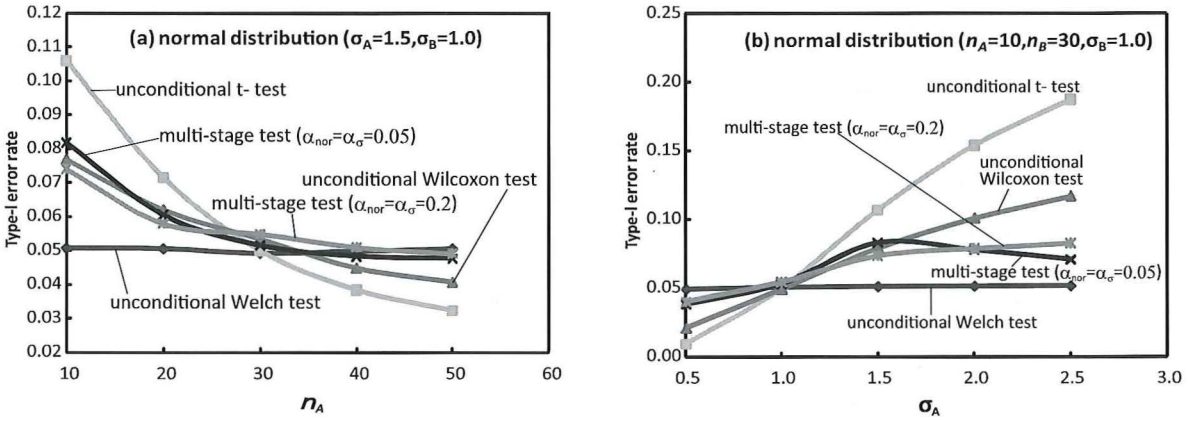

図３　第１種の過誤の発生率シミュレーション結果。(a)は、標本サイズの影響を調べる目的で、$\sigma_A=1.5$, $\sigma_B=1.0$, $\mu_A=\mu_B=0.0$, $n_B=30$ に固定して、第1群標本サイズだけを変化させた。(b)は標準偏差の影響を調べる目的で, $\sigma_B=1.0$, $\mu_A=\mu_B=0.0$, $n_A=10$, $n_B=30$ に固定して、第1群標準偏差だけを変化させた。多段階検定の結果については、$\alpha_{nor}=\alpha_\sigma=0.05$ に設定した場合と、$\alpha_{nor}=\alpha_\sigma=0.2$ に設定した場合の両方の結果を示した。

本サイズは30に固定した。

図３の結果より、以下が言える。

＊　多段階検定の第１種の過誤の発生率は、標本サイズが等しい場合（図３(a)で$n_B=30$の場合）および等分散の場合（図３(b)で$\sigma_A=1.0$の場合）を除いて、理論的な理想値である 0.05 からずれ、標本サイズの不均一性および分散の不均一性が大きくなるとそのずれが顕著となる。

＊　上記多段階検定の第１種の過誤の発生率の挙動は、正規性の検定および等分散検定の有意水準をそれぞれ$\alpha_{nor}=\alpha_\sigma=0.05$に設定した場合および$\alpha_{nor}=\alpha_\sigma=0.2$に設定した場合で、ほとんど同じである.

＊　第１種の過誤の発生率が設定した有意水準である$\alpha_\mu=0.05$に最も近いのは、無条件 Welch 検定を適用した場合である。無条件 Welch 検定を適用した場合、両群のサンプルサイズの不均一および母分散の不均一の影響をほとんど受けることなく、第１種の過誤の発生率は安定して設定した有意水準に非常に近くなる。

＊　サンプルサイズの不均一および母分散の不均一の影響の度合いから、使用した検定方式の安定性が優れている順に順位を付けると、

1　無条件 Welch 検定
2　多段階検定
3　無条件 Wilcoxon 検定
4　無条件 t-検定

の順になる。ただし、多段階検定と無条件 Wilcoxon 検定との安定性の違いはごくわずかである。

考　察

１．各検定方式の安定性について

まず、無条件 Welch 検定が最も優れている理由について考察する。Welch 検定の適用条件は、表１に示した通り両群の母集団分布が正規分布に従っていることであり、この条件は今回のシミュレーションにおいては完全に満たされている。従って、無条件 Welch 検定の第１種の過誤の発生率が理論値である0.05に近い値を示すことは、納得ができる結果と言える。

続いて、多段階検定を除く各検定方式の安定性について考察する。t-検定の適用条件は母集団の正規性および母分散が等しいことであるが、シミュレーション設定でこの条件が完全に満たされるのは分散を等しく設定した場合（図３(b)で$\sigma_A=1.0$とした場合）のみである。これ以外のシミュレーション設定条件はすべて前提を満たしていないため、無条件 t-検定が最も安定性を欠く結果になったと思われる。これに対して、Wilcoxon 検定は順位和を使用する正規性を前提としない検定方式であるため、一般に外れ値の影響を受けにくく、t-検定よりも頑健であるとされている。図３で、無条件 Wilcoxon 検定が t-検定より安定した挙動を示しているという結果はこれと符合する結果である。

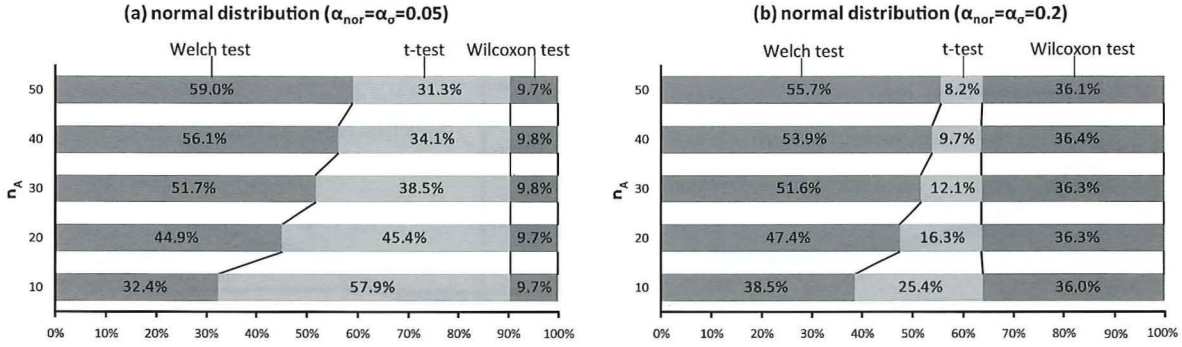

図４　多段階検定における２群の位置に関する検定のためにそれぞれの検定手法が選択される割合。$\sigma_A=1.5$, $\sigma_B=1.0$, $\mu_A=\mu_B=0.0$, $n_B=30$ に固定して、第１群標本サイズだけを変化させた。(a)は、$\alpha_{nor}=\alpha_\sigma=0.05$ に設定した場合を、(b)は $\alpha_{nor}=\alpha_\sigma=0.2$ に設定した場合の割合を示している。

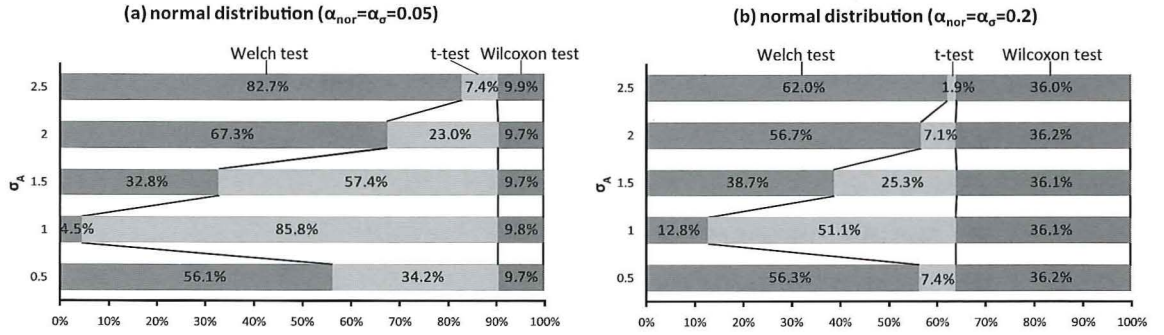

図５　多段階検定における２群の位置に関する検定のためにそれぞれの検定手法が選択される割合。$n_A=10$, $n_B=30$, $\sigma_A=1.5$, $\mu_A=\mu_B=0.0$ に固定して、第1群標準偏差だけを変化させた。(a)は、$\alpha_{nor}=\alpha_\sigma=0.05$ に設定した場合を、(b)は $\alpha_{nor}=\alpha_\sigma=0.2$ に設定した場合の割合を示している。

２．多段階検定における各検定方式が選択される割合

多段階検定の安定性について考察するためには、第１種の過誤の発生率のシミュレーション実行時に２標本の位置に関する検定手法として Welch 検定、t-検定、Wilcoxon 検定のそれぞれがどれくらいの割合で選択されたのかの情報が重要である。第１群の標本サイズおよび標準偏差設定値を変化させることにより、正規性の検定および等分散検定で帰無仮説が棄却される割合が異なり、その結果として最終的に選択される２標本の位置の検定方式の割合も変化する。図４、５に、多段階検定で第１群の標本サイズおよび標準偏差を変化させた各条件において、選択された２群の位置に関する検定方式の割合を示す。

図４(a)は、図３(a)と同じ設定条件（$\sigma_A=1.5$、$\sigma_B=1.0$、$n_B=30$、n_A を変化させる）で $\alpha_{nor}=\alpha_\sigma=0.05$ に設定した場合の各検定方式が選択される割合を、図４(b)は $\alpha_{nor}=\alpha_\sigma=0.2$ に設定した場合の割合を示している。図４(a)、(b)について、以下の解釈が可能である。

＊ Wilcoxon 検定が選択される割合は、図４(a)では 9.7～9.8%で一定、図４(b)では 36.0～36.4%でこちらもほぼ一定である。この結果から、正規性の検定（Shapiro-Wilk 検定）が非常に信頼性の高いものであることが伺える。図２の枠組みを参照すれば、Wilcoxon 検定は正規性の検定で帰無仮説（母集団は正規分布に従っている）がいずれかの群または両群ともに棄却されたときに選択されるが、少なくともいずれか一方の群について帰無仮説が棄却される確率は

$$P(A\cup B)=P(A)+P(B)-P(A\cap B) \quad \cdots\cdots(1)$$

で与えられ、この値は $P(A\cup B)=0.0975$（$\alpha_{nor}=0.05$ のとき）または $P(A\cup B)=0.36$（$\alpha_{nor}=0.2$ のとき）になるからである。ただし、正規性の検定における第１群に関する棄却域に対応する事象を A、第２群に関する棄却域に対応する事象を B と表記した。

＊ 図４(a)、(b)いずれにおいても、t-検定が選択される割合は、第１群の標本サイズが増加するにつれて減少している。t-検定は前述のように等

分散検定が棄却された場合に選択される。ここでの設定は$\sigma_A=1.5$、$\sigma_B=1.0$であり、等分散ではない．従って、図４(a)、(b)の結果は、第１群の標本サイズが多くなるにつれて等分散検定の帰無仮説が正しく棄却される割合が増加することを示している。

＊　t-検定が選択される割合は、図４(a)が大きく図４(b)では少ない。図４(a)は$\alpha_\sigma=0.05$、図４(b)では$\alpha_\sigma=0.2$であるから、等分散検定の棄却域は図４(b)の場合のほうが大きいため、これも当然の結果であると言える。

図５(a)は、図３(b)と同じ設定条件（$n_A=10$、$n_B=30$、$\sigma_B=1.0$、σ_Aを変化させる）で$\alpha_{nor}=\alpha_\sigma=0.05$に設定した場合の各検定方式が選択される割合を、図５(b)は$\alpha_{nor}=\alpha_\sigma=0.2$に設定した場合の各検定方式が選択される割合を示している。図５(a)、(b)について、以下の解釈が可能である。

＊　Wilcoxon 検定が選択される割合は、図４(a)では9.7～9.9%で一定、図４(b)では36.0～36.2%でこちらも一定である。この結果は図４(a)、(b)と同一であり、ここでも正規性の検定が非常に信頼性の高いものであることが分かる。すなわち、正規性の検定は、標本サイズの不均一性および分散の不均一性の影響をほとんど受けずに、第１種の過誤の発生率を理論通りの値（有意水準）に制御することができる。

＊　図５(a)、(b)いずれにおいても、t-検定が選択される割合は$\sigma_A=1.0$の場合が最も多く、σ_Aの値がこれより大きくても小さくてもその割合は減少する。$\sigma_A=1.0$の場合は等分散であり、それ以外の場合は不等分散であるから、この割合の変化は当然の結果であると言える。

＊　t-検定が選択される割合は、図５(a)が大きく図５(b)では少ない。図５(a)は$\alpha_\sigma=0.05$、図５(b)では$\alpha_\sigma=0.2$であるから、等分散検定の棄却域は図５(b)の場合のほうが大きいため、これも当然の結果であると言える。

３．多段階検定における各検定方式の第１種の過誤の発生率

図４、５では、多段階検定において Welch 検定/t-検定/Wilcoxon 検定がそれぞれどのような割合で選択されるのかを見たが、ここでは、これらの検定方式が選択された場合に、それぞれの第１種の過誤の発生率がどのような挙動を示すかを調べる。多段階検定における各検定方式の第１種の過誤の発生率を、次式で定義する。

$$(\text{第1種の過誤の発生率})=\frac{\text{各検定方式で帰無仮説が棄却された回数}}{\text{各検定方式が選択された回数}} \quad (2)$$

図６(a)、(b)に、多段階検定における各検定方式の棄却率の変化を示す。図６(a)は第１群の標本サイズを変化させた場合、図６(b)は第１群の分散を変化させた場合であり、ともに正規性の検定および等分散検定の有意水準を$\alpha_{nor}=\alpha_\sigma=0.05$に設定した場合を示した。($\alpha_{nor}=\alpha_\sigma=0.2$に設定した場合はここでは示さないが、各検定方式の第１種の過誤の発生率の挙動は0.05の場合とほぼ同一である。)

図６(a)では、Wilcoxon 検定およびt-検定の第１種の過誤の発生率はn_Aの小さい領域で0.1を超える大きな値を示し、n_Aが大きくなるにつれて小さくなる。今回の設定では、Wilcoxon 検定が選択されるのは母集団分布が左右非対称な分布と見なされるような標本が使用されるときであると考えられるから、このような標本に対して Wilcoxon 検定において帰無仮説が棄却される確率すなわち第１種の過誤の発生率は理想的な値である0.05より大きくなり、このずれが標本サイズn_Aの増加に伴って改善されているものと解釈できる。一方、t-検定が選択される場合として、たとえば第１群の母集団分布の標準偏差が第２群と同じ値である$\sigma_A=1.0$と判断されるような標本が多く含まれていると考えられるが、この標準偏差は本来の設定値である$\sigma_A=1.5$より小さいため、この場合の標本は偏った（本来の母集団を正しく表していない）標本になっているはずである。この標本には、その平均値が母平均と一致するものも含まれるが、そうではない（大小いずれかの方向に偏った）標本も含まれるため、t-検定において帰無仮説が棄却される確率(すなわち第１種の過誤の発生率)は0.05より高くなり、このずれが標本サイズn_Aの増加に伴って改善されているものと解釈できる。

一方図６(a)における Welch 検定の第１種の過誤の発生率は、n_Aの小さい領域では理想値である0.05と

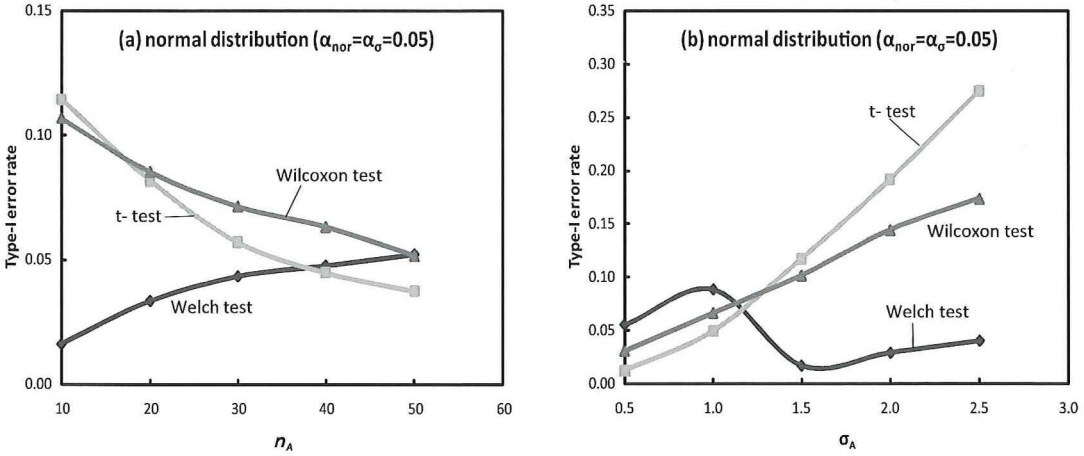

図６　多段階検定における２群の位置に関する検定のための各検定手法が示す第1種の過誤の発生率。(a)は $\sigma_A=1.5, \sigma_B=1.0, \mu_A=\mu_B=0.0, n_B=30$ に固定して、第1群標本サイズだけを変化させた場合を、(b)は、$n_A=10$, $n_B=30, \sigma_B=1.0, \mu_A=\mu_B=0.0$ に固定して、第１群標準偏差だけを変化させた場合を示す。(a),(b)いずれも $\alpha_{nor}=\alpha_\sigma=0.05$ に設定した場合の割合を示した。

比較するとかなり小さく、n_Aの増加につれて0.05に近づく。n_Aの小さい領域では、無条件 Welch 検定においては帰無仮説の棄却の対象となっていたような左右非対称な標本や偏った標本の多くは、多段階検定においては正規性の検定および等分散検定の際に棄却されて、それぞれ Wilcoxon 検定および t-検定に流れるという状況が発生していると考えられる。そのため、n_Aの小さい領域で Welch 検定が対象としている標本の多くは、本来の母集団の素性を正しく反映している標本であり、Welch 検定における第１種の過誤の発生率はこのために0.05より小さくなると考えられる。

図６(b)では、Wilcoxon 検定および t-検定の第１種の過誤の発生率は、σ_Aの増加に伴って増加している。σ_Aの大きい領域で Wilcoxon 検定が対象とする第１群の標本は、正規性の検定で棄却された左右非対称な標本であり、それが極端な場合には当然 Wilcoxon 検定における第１種の過誤が発生する確率が高くなる。一方σ_Aの大きい領域で t-検定が対象とする標本は、第１群の母集団分布の標準偏差が第２群と同じ値である$\sigma_A=1.0$と判断されるような標本であるから、実際のσ_Aより標準偏差が小さいと判断されるような偏った標本である。この中には、平均値が本来の母集団の平均値と一致しないものが相当数含まれているはずであり、これがσ_Aが大きい領域でt-検定の第１種の過誤の発生率が高くなる理由であると考えられる。

また、図６(b)では Welch 検定の第１種の過誤の発生率が$\sigma_A=1.0$の点で例外的に大きくなっている。この点においては、母集団分布の設定は$\sigma_A=\sigma_B=1.0$であるため、等分散検定の帰無仮説が採択される確率が高くなり、これらの標本はすべて t-検定に流れる。したがって、$\sigma_A=1.0$の点で Welch 検定が対象とする標本は、等分散である母集団分布から偏って等分散とは見なさないような標本化をされた標本だけであり、その結果第１種の過誤の発生率が増加しているものと考えられる。

４．多段階検定の妥当性について

これまでに述べてきたシミュレーション結果から、図２に示した多段階検定の妥当性について考察する。

多段階検定では、パラメトリック検定/ノンパラメトリック検定のいずれを適用するかを判断するために、まず正規性の検定を行う。今回シミュレーションで使用した正規性の検定手法である Shapiro-Wilk 検定そのものは、母集団が正規分布である場合に第１種の過誤の発生率をほぼ有意水準と一致する値にコントロールできる、非常に信頼性の高い検定であると判断できることはすでに述べた。しかし、この検定を多段階検定に使用する場合には、以下のような問題が生じる。

＊　パラメトリック検定(t-検定または Welch 検定)を適用することの妥当性を主張するために、$\alpha_{nor}=0.2$を用いて正規性の検定を行うケース

を考える。この場合の検定結果として、両群とも正規分布に従う/少なくともどちらか一方は正規分布に従わない、のどちらになるかは完全に標本のサンプリングのされ方に依存しており、その比率は標本サイズに依存せずに約６:４となる（図４(b)および図５(b)参照のこと）。この結果は、正規性の検定が統計手法選択のための基準とはならないことを意味している。

＊　逆に、ノンパラメトリック検定（Wilcoxon 検定）を適用することの妥当性を主張するため $\alpha_{nor}=0.05$ を用いて正規性の検定を行うケースを考える。この場合は、約１割は誤ってノンパラメトリック検定に流れることになる（図４(a)および図５(a)参照のこと）。このケースでも、正規性の検定が統計手法選択の指針とはならないことが分かる。

さらに、正規性の検定および等分散検定を最終的な２標本の位置の検定の前に行うことで、各検定方式に供給される標本が偏った性質をもつようになることも、これまでの議論から明らかである。すなわち、Wilcoxon 検定には左右非対称な標本が数多く提供され、また t-検定には等分散と見なされるような標本のみが、また Welch 検定には不等分散と見なされるような標本のみが供給される。こうした偏った構成の標本が選択的に各検定手法に供給される結果、多段階検定においては本来各検定手法が持っていたはずの第１種の過誤の発生率の安定性が著しく損なわれる結果になる。これは、図３(a)と図６(a)、図３(b)と図６(b)のそれぞれで、無条件に各検定手法を適用した場合の第１種の過誤の発生率の挙動と、多段階検定における各検定手法の第１種の過誤の発生率の挙動を比較すれば明らかである。上記のそれぞれの比較において、無条件にその手法を適用したほうが多段階検定における挙動よりも安定して第１種の過誤の発生率が0.05により近くなることが分かる。

結　論

シミュレーション結果およびその考察から判断すると、現在広く用いられている多段階検定の枠組みは、論理的に妥当性を欠くものであると言わざるを得ない。本研究で得られた主要な結果を以下に列挙する。

１　多段階検定は、適切な検定手法を選択するための指針として機能しない。

２　多段階検定により得られた検定結果は、第１種の過誤の発生率をコントロールできていないという観点から、信頼性に欠けるものである。

３　多段階検定が妥当性を欠く枠組みであるという上記結論は、正規性の検定および等分散検定における有意水準の設定値の大小に関わらず言えることである。

４　シミュレーションで使用した各検定方式の中では、無条件で Welch 検定を適用する方法が最も妥当なものである。この場合は、母集団分布の正規性および等分散性の成立の程度とは無関係に、第１種の過誤の発生率はほぼ有意水準と同じ値を示す。すなわち、第１種の過誤の発生率を理論値通りにコントロールすることが可能である。

本研究では、図２の枠組みについてのシミュレーションを行った。先行研究として、等分散検定の影響を調べた事例はすでに報告されているが[15]、正規性の検定も含めてその影響を調べた研究は本研究が初めてであると思われる。今後は本研究では対象外とした図１の枠組みについても研究を進める予定である。また、これ以外の枠組みとして、正規性の検定の結果により Welch 検定および Brunner-Munzel 検定の２つの検定手法のいずれかに振り分けるという枠組みも考えられる。この場合の妥当性の検証も、今後の課題としたい。

謝　辞

本論文をまとめるにあたり，有益な議論をしていただいた樋口雄介氏に感謝いたします.

文　献

1)　Wilcox RR: Comparing the means of two independent groups, Biometrical Journal, 32, 771-780, 1990.

2)　Wilcox RR, Keselman HJ: Modern robust data analysis methods:measures of central tendency, *Psychological Methods*, 8, 254-274, 2003.

3) 天野弘美他:２標本の位置の検定に関するコンピュータ・シミュレーション, 昭和大学富士吉田教育部紀要, 第４巻, 17-28, 2009.

4) Eilertsen AL, Qvigstad E, Andersen TO et al: Conventional-dose hormone therapy (HT) and tibolone, but not lowdoseHT and raloxifene, increase markers of activated coagulation, Maturitas, 55, 278-87, 2006.

5) Morten W. Fagerland , Leiv Sandvik: Performance of five two-sample location tests for skewed distributions with unequal variances, Contemporary Clinical Trials, 30, 490-496, 2009.

6) Skovlund E, Fenstad GU: Should we always choose a nonparametric test when comparing two apparently nonnormal distributions?, Journal of Clinical Epidemiology, 54, 86-92, 2001

7) Markus Neuhauser: A nonparametric two-sample comparison for skewed data with unequal variances, Journal of Clinical Epidemiology, 63, 691-693, 2010

8) Student : The Probable error of a mean , Biometrika,6, 1-25,1908.

9) Welch BL:The significance of the difference between two means when the population variances are unequal, Biometrika, 29, 350-62, 1937.

10) Wilcoxon F:Individual comparisons by ranking methods, Biometrics Bulletin, 1, 80-83,1945.

11) Mann H. B., and Whitney D. R.:On a test of whether one of two random variables is stochastically larger than the other, Annals of Mathematical Statistics,18,50-60, 1947.

12) Brunner E, Munzel U.:The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation, Biometrical Journal,42, 17-25, 2000.

13) Reiczigel J., Zakarias I, and Rozsa L.:A bootstarp test of stochastic equality of two populations, The American Statistician, 6, 1-6, 2005.

14) R Development Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

15) Zimmerman, D. W.:A note on preliminary tests of equality of variances, British Journal of Mathematical & Statistical Psychology, 57, 173-181, 2004.

A computer simulation study on the performance of two-sample location tests (II)

Hiromi AMANO[1) 2)], Masato KONDOU[2)],
Hiroshi OGURA[2)], Toshikazu TAKAGI[2)]

[1)] School of Nursing and Rehabilitation Sciences, Showa Unversity
[2)] Faculty of Arts and Sciences at Fujiyoshida, Showa Unversity

Abstract

Comparison of locations, or central tendency, of two independent populations is common in various statistical researches. To select an appropriate hypothesis test for comparing locations among several available methods, the preliminary test of normality and that of equality of variances are widely used before conducting a test of location. In this study, we performed simulation experiments of a multi-stage procedure including the preliminary tests in order to confirm the validity of such multi-stage tests for comparing two-sample locations. Effects of sample size and variance heterogeneities on the test performance were examined. Results of simulations show that a inflation of Type-I error rates occurs when the multi-stage procedure is applied, which indicates that the preliminary tests usually make situations worse and the multi-stage procedure is no longer recommended. On the other hand, the unconditional use of the Welch test without preliminary steps shows stable Type-I error rates that are very close to the predefined significance level of 5%.

Key Words: two-sample location test, preliminary test, t-test, Welch test, Wilcoxon test, Brunner-Munzel test